

Estimating Population Sizes with Link-Tracing Sampling

Kyle Vincent, Steve Thompson

Department of Statistics and Actuarial Science

Simon Fraser University, 8888 University Drive

Burnaby, British Columbia, CANADA

V5A 1S6

email: kvincent@sfu.ca, thompson@sfu.ca

October 25, 2012

Abstract

We solve a problem in the capture-recapture literature which involves estimating the size of networked hard-to-reach populations with independent samples that are selected with a link-tracing design. Our novel method introduces an advantage over typical capture-recapture studies as the inference procedure is based on a sufficient statistic which in turn allows for adaptively recruited members of the target population to be included in the analysis. Preliminary capture-recapture style estimates of the population size are based on randomly selected initial samples. Rao-Blackwellization of the preliminary estimator entails averaging over preliminary estimates obtained from the full sample reorderings that are consistent with a sufficient statistic which incorporates information from the adaptively recruited members. We evaluate the new inferential method for two link-tracing designs applied to a simulated networked population. The results from this two-sample study demonstrate that our inferential method will provide more efficient estimates of the population size relative to the estimators found in the existing capture-recapture literature.

Keywords: Adaptive sampling; Capture-recapture; Design-based inference; Lincoln-Petersen estimator; Link-tracing designs; Population size estimates; Rao-Blackwellization.

1 Introduction

In this article we introduce a new design-based method for estimating the size of networked hard-to-reach populations based on independent samples when selected through a link-tracing design. Our novel method now permits for adaptively selected members of the target population to be included in the inference procedure through a Rao-Blackwellization method based on a sufficient statistic. Moreover, our method possesses an additional advantage over the existing inferential methods based on estimating population sizes with link-tracing designs; our method permits for the possibility of obtaining members not within arms reach of the initial sample (that is, those not immediately linked to the initial sample).

There has been a growing interest in the use of statistical methods for estimating characteristics of hard-to-reach populations like those comprised of injection drug-users, commercial sex-workers, and the homeless. Such hard-to-reach populations may have a tendency to exhibit social links between members based on a predefined relationship like the sharing of drug-using paraphernalia or coming into sexual contact. Since such hard-to-reach populations consist of individuals that may be extremely difficult to locate and recruit for research study purposes, these social links may be used by a researcher to adaptively recruit more members of the target population, thereby increasing the sample size for efficiency gains of estimators when inferring on the population unknowns.

There is a growing body of literature on both model-based and design-based approaches to making inference for population characteristics through the use of adap-

tive sampling strategies when the population size is assumed known. Thompson and Frank (2000) described an approach to likelihood-based inference for link-tracing sampling designs. This approach was used in further development in Chow and Thompson (2003), and Handcock and Giles (2010) developed a theoretical framework for basing inference of population unknowns on exponential random graph models when using an adaptive sampling design. Thompson (2006) generalized the design-based method for estimating population proportions based on an adaptive sampling strategy termed adaptive web sampling that allows for fixed sample sizes as well as the flexibility to allocate as much random or adaptive effort as desired at each step in the sample selection procedure. For additional information, Feinberg (2010a and 2010b) provides a summary and discussion of some of the work on the modeling and analysis of networked populations, as well as a general introduction to papers with applications towards sampling and analyzing rare and social populations. Spreen (1992) also provides some review of link-tracing designs and their applicability for sampling from hard-to-reach populations.

Heckathorn (1997 and 2002) developed a procedure termed respondent-driven sampling which bases estimates of population proportions on Markov chain theory. Abdul-Quader et al. (2006) describes empirical findings based on a respondent-driven sampling design to collect data on a HIV-related population in the New York City area. Recent work by Gile and Handcock (2011) proposes a modified estimator of population means when employing respondent-driven sampling that utilizes a model-assisted approach to help overcome the initial bias introduced with the selection of an initial sample that is not a probability sample. As their inference procedure requires

knowing the population size, they show that this new estimator is robust towards when the inference procedure substitutes a relatively large or small estimated value for the true population size.

There are many methods for estimating population sizes through a capture-recapture style of study (see Schwarz and Seber (1999), and Chao et al. (2001) for a summary of the existing methods), and some of these classic methods have been implemented for estimating the size of hidden drug-using populations (see Frischer et al. (1993) and Mastro et al. (1994)). Several model-based approaches for estimating population sizes with the use of a link-tracing design have been developed for when a subset of the target population is accessible from a sampling frame. Felix-Medina and Thompson (2004) developed an approach that combines model-based and design-based inference. It assumes that links from the partial sampling frame are made with a homogenous pattern that facilitates a capture-recapture style of inference. Felix-Medina and Monjardin (2006) extend on this work by proposing a Bayesian-assisted approach to overcome some of the bias that the maximum likelihood estimators present. Many hidden populations have a high degree of unpredictable behavior, for example in the form of erratic clustering patterns amongst its members. In such situations model-based estimators may not be a robust measure for the population size because departures from the assumptions that give rise to such an analysis are occurring.

Frank and Snijders (1994) developed a design-based approach to inference that formulates consistent moment-based estimates of the population size based on links recorded within and outside of a Bernoulli sample. However, the design is limited to

only observing links from members in the Bernoulli sample.

The method outlined in this article consists of selecting initial samples at random and then adaptively tracing links out of each current sample according to a predetermined probability sampling mechanism. Adaptive sampling designs have a tendency to yield members of the population with a larger degree (ie. number of neighbors) relative to a completely random sample. In consequence, classic capture-recapture estimators like the Lincoln-Petersen estimator will likely underestimate the population size. To correct for this, we use the capture-recapture style estimator to use information from the initial random samples and then exploit a sufficient statistic to incorporate the information from adaptively recruited members. Our method averages the capture-recapture estimates using a sufficient statistic based on hypothetical reorderings of the independent samples that are consistent with the sufficient statistic. We use a simulation study to show that the additional sampling efforts required for adaptive recruitment will result in a significant gain in precision over the preliminary estimates.

In Section 2 we introduce the notation that is used in this article. In Section 3 we outline the link-tracing sampling designs that are explored, namely one which is analogous to the general adaptive web sampling design that was introduced by Thompson (2006) as well as a nearest neighbors adaptive web sampling design that has the potential for more practical use for sampling from an empirical hidden population. Section 4 is reserved for developing estimates of the population size and average node degree of the population as well as the variances of these estimates. As tabulating the preliminary estimates from all reorderings of the final samples

is computationally cumbersome for the samples selected in this study, in Section 5 we outline a Markov chain resampling procedure to obtain estimates of the Rao-Blackwellized estimates. In Section 6 we perform a two-sample simulation study for the two sampling designs on a simulated networked population, and then draw conclusions and provide a general discussion of the novel methods developed in this article in Section 7.

2 Sampling Setup

We define a population U to consist of the set of units/individuals $U = \{1, 2, \dots, N\}$ where N is the population size. Each pair of units (i, j) , $i, j = 1, 2, \dots, N$, is associated with a weight w_{ij} . In this study we set $w_{ij} = 1$ if there is a link (or predetermined relationship) from unit i to unit j , and zero otherwise. We define $w_{ii} = 0$ for all $i = 1, 2, \dots, N$ in the population.

Our sampling approach consists of two stages. First we select K independent initial random samples, and from these samples we trace social links out of the current samples, based on a predetermined probability sampling mechanism, independently between and without replacement within each sample. Each sample is selected based on a design that is equivalent to an adaptive web sampling design that is selected without allowing for random jumps after the initial samples are selected (Thompson, 2006). The observed data is $D_0 = \{(i, t_{i,k}, w_{ij}, w_i^+, k) : i, j \in s_k, k = 1, 2, \dots, K\}$ where s_k refers to sample k for $k = 1, 2, \dots, K$; $t_{i,k}$ is the time (or step) in the sampling sequence that unit i is selected for sample k ; w_i^+ is the out-degree of

individual i (ie. the number of members acknowledged by individual i). A sufficient statistic based on the full data set is $T_S = \{(i, w_{ij}, w_i^+, k) : i, j \in s_k, k = 1, 2, \dots, K\}$. This statistic is sufficient, but T_S is not the minimal sufficient statistic (as described by Thompson and Seber (1996)). The minimal sufficient statistic would consist of reduced data from both samples. In this case the minimal sufficient statistic would be $T_M = \{(i, w_{ij}, w_i^+) : i, j \in s\}$ where $s = \bigcup_{k=1}^K s_k$. Using T_M would entail tabulating preliminary estimates from all reorderings of the combined samples. Calculating such an estimate is not computationally feasible due to extensive resampling requirements and for this reason we do not utilize the minimal sufficient statistic in our study.

3 The Sampling Designs

Thompson (2006) outlined the general method for selecting an adaptive web sample without replacement. In this article, we explore the use of two different adaptive web sampling designs, the first being the general design outlined by Thompson (2006) and the second being a nearest neighbors adaptive web sampling design. In this section, we will outline the selection process for the two adaptive web sampling designs.

The first sampling design selects independently two adaptive web samples as follows. The sampling procedure commences with the selection of an initial sample s_0 of size n_0 of members from the population completely at random. A predetermined maximum number of individuals, $n - n_0$ say, are further selected sequentially to bring the sample size up to $n' \leq n$ by tracing links, when available, out of the current active set as follows. For any step j , $j = 1, 2, \dots, n - n_0$, any member i that

has not yet been selected and is linked to at least one member in the current active set is selected for inclusion with probability $q_{j,i} = \frac{w_{a_j,i}}{w_{a_j,+}}$, where $w_{a_j,i}$ is the number of links from the current active set a_j to unit i , and $w_{a_j,+}$ is the number of links out of the current active set to members not yet selected. Hence, if the number of links to trace out of the current active set is exhausted at any intermediate step in the sampling process then sampling stops at the most recent step $j - 1$ so that the final sample size is $n' = n_0 + j - 1$. In the general adaptive web sampling design, the active set consists of all members that have been selected for the current sample so that recruitment at any intermediate step is based on all links stemming out of the current sample. That is, for any step j , $j = 1, 2, \dots, n - n_0$, any member i not yet chosen is selected with probability $q_{j,i} = \frac{w_{s_j,i}}{w_{s_j,+}}$ where s_j represents the current sample. Notice that with the general adaptive web sampling design it is possible to select members that are not directly linked to the initial sample.

The nearest neighbors adaptive web sampling design restricts the active set to consist only of those members who are selected for the initial sample so that only the units that are linked to the initial sample have a positive probability of being selected for the final sample. To clarify, for any step j , $j = 1, 2, \dots, n - n_0$, any member i not yet chosen is selected with probability $q_{j,i} = \frac{w_{s_{0,j},i}}{w_{s_{0,j},+}}$ where $w_{s_{0,j},i}$ is the number of links from the initial sample out to unit i at step j and $w_{s_{0,j},+}$ is the number of links out of the initial sample to members not yet selected at step j .

For $k = 1, 2, \dots, K$ we shall let s_{0k} represent the initial random sample corresponding with sample k where $|s_{0k}| = n_{0k}$. We shall also let s_k represent the final sample k in the order it was selected, where $|s_k| = n'_k = n_{01}, n_{01} + 1, \dots, n_k$. For inferential

purposes we shall define $s_{(1,2,\dots,K)}$ to be the full ordered sample of the samples in the respective order they were selected. The probability of selecting any sample $s_{(1,2,\dots,K)}$ can then be expressed as

$$p(s_{(1,2,\dots,K)}) = \prod_{k=1}^K \left(\frac{1}{\binom{N}{n_{0k}}} \prod_{t_k=0}^{n'_k - n_{0k}} q_{t_k}^{s_k} \right).$$

The first terms in the expression correspond with the random selection of the initial samples and $q_{t_k}^{s_k}$ is the probability of adaptively selecting the unit that was selected at step t_k for sample k . It shall be understood that for $t_k = 0$, $q_{t_k}^{s_k} = 1$ for $k = 1, 2, \dots, K$.

4 Estimation

4.1 Population size estimators

Suppose that \hat{N}_0 is a preliminary estimate of the population size based on the K initial random samples. An improved estimator based on the aforementioned sufficient statistic T_S is

$$\hat{N}_{RB} = \sum_{r_1=1}^{n'_1!} \sum_{r_2=1}^{n'_2!} \cdots \sum_{r_K=1}^{n'_K!} \hat{N}_0^{(r_1, r_2, \dots, r_K)} p(s_{(r_1, r_2, \dots, r_K)} | T_S)$$

where $\hat{N}_0^{(r_1, r_2, \dots, r_K)}$ is the preliminary population size estimate based on the hypothetical initial samples corresponding with reorderings r_1, r_2, \dots, r_K of samples $1, 2, \dots, K$,

respectively, and $p(s_{(r_1, r_2, \dots, r_K)} | T_S)$ is the conditional probability of obtaining the sample reorderings r_1, r_2, \dots, r_K given the data observed for T_S . Notice that for any specific sample reordering $s_{(x_1, x_2, \dots, x_K)}$ (that is, x_1, x_2, \dots, x_K are specific indices of the countable permutations of s_1, s_2, \dots, s_K , respectively), the conditional probability of obtaining the sample reordering in this respective order can be expressed as

$$\begin{aligned}
p(s_{(x_1, x_2, \dots, x_K)} | T_S) &= p(s_{(x_1, x_2, \dots, x_K)}) / \sum_{r_1=1}^{n'_1!} \sum_{r_2=1}^{n'_2!} \cdots \sum_{r_K=1}^{n'_K!} p(s_{(r_1, r_2, \dots, r_K)}) \\
&= \frac{1}{\binom{N}{n_{01}}} \prod_{t_1=0}^{n'_1-n_{01}} q_{t_1}^{s_{x_1}} \times \frac{1}{\binom{N}{n_{02}}} \prod_{t_2=0}^{n'_2-n_{02}} q_{t_2}^{s_{x_2}} \times \cdots \times \frac{1}{\binom{N}{n_{0K}}} \prod_{t_K=0}^{n'_K-n_{0K}} q_{t_K}^{s_{x_K}} / \\
&\quad \sum_{r_1=1}^{n'_1!} \sum_{r_2=1}^{n'_2!} \cdots \sum_{r_K=1}^{n'_K!} \left(\frac{1}{\binom{N}{n_{01}}} \prod_{t_1=0}^{n'_1-n_{01}} q_{t_1}^{s_{r_1}} \times \frac{1}{\binom{N}{n_{02}}} \prod_{t_2=0}^{n'_2-n_{02}} q_{t_2}^{s_{r_2}} \times \cdots \times \frac{1}{\binom{N}{n_{0K}}} \prod_{t_K=0}^{n'_K-n_{0K}} q_{t_K}^{s_{r_K}} \right) \\
&= \prod_{t_1=0}^{n'_1-n_{01}} q_{t_1}^{s_{x_1}} \times \prod_{t_2=0}^{n'_2-n_{02}} q_{t_2}^{s_{x_2}} \times \cdots \times \prod_{t_K=0}^{n'_K-n_{0K}} q_{t_K}^{s_{x_K}} / \\
&\quad \sum_{r_1=1}^{n'_1!} \sum_{r_2=1}^{n'_2!} \cdots \sum_{r_K=1}^{n'_K!} \left(\prod_{t_1=0}^{n'_1-n_{01}} q_{t_1}^{s_{r_1}} \times \prod_{t_2=0}^{n'_2-n_{02}} q_{t_2}^{s_{r_2}} \times \cdots \times \prod_{t_K=0}^{n'_K-n_{0K}} q_{t_K}^{s_{r_K}} \right).
\end{aligned}$$

Notice that all terms involving the unknown population size N are factored out of the expressions and can be canceled to make computation of the Rao-Blackwellized estimates possible. This may not be the case if random jumps were permitted after selecting the initial samples in the sampling design as these new selection probabilities would depend on N in such a manner that factoring N out of the expression may not be possible.

In our two sample study, a preliminary estimate of the population size based on the initial random samples is the Lincoln-Petersen estimator (Petersen (1896)). As this estimator can result in unstable estimates when there is no recorded overlap between the two initial random samples we will use the bias-adjusted Lincoln-Petersen (LP) estimator proposed by Chapman (1951). This estimator is of the form

$$\hat{N}_0 = \frac{(n_{01} + 1)(n_{02} + 1)}{m + 1} - 1, \quad (1)$$

where m denotes the number of individuals that are selected for both initial samples s_{01} and s_{02} .

4.2 Alternative population size estimator for the two-sample study

An estimator that improves on the preliminary estimator found in (1) is obtained by constructing the Lincoln-Petersen estimator based on the initial random selection for the first sample and the full sample for the second sample. This estimator can be expressed as

$$\hat{N}_{0,1} = \frac{(n_{01} + 1)(n'_2 + 1)}{m_1 + 1} - 1,$$

where n_{01} is the initial sample size of the first sample, n'_2 is the full sample size

of the second sample, and m_1 is the number of individuals selected for both initial sample 1 and final sample 2. Notice that this estimator is more efficient than the aforementioned preliminary estimator since all of sample 2 is utilized at the inference stage. The Rao-Blackwellized version of this estimator can also be shown to not depend on the unknown population size N and can be expressed as

$$\hat{N}_{RB,1} = \sum_{r_1=1}^{n'_1!} \hat{N}_{0,1}^{(r_1,s_2)} p(s_{(r_1)}|T_{S_1}),$$

where $\hat{N}_{0,1}^{(r_1,s_2)}$ is the estimate of N obtained with the hypothetical initial sample of reordering r_1 of sample 1 and all of sample 2 (notice that this estimate does not depend on the ordering of sample 2), $p(s_{(r_1)})$ is the probability of obtaining reordering r_1 of sample 1, and $T_{S_1} = \{(i, w_{ij}, w_{i+}) : i, j \in s_1\} \cup \{i : i \in s_2\}$. Notice that T_{S_1} is a function of T_S , and hence the estimator $\hat{N}_{RB,1}$ will result in a more efficient estimate of the population size relative to \hat{N}_{RB} .

We shall note here that in the event that a fixed list of n'_2 individuals from the target population has been previously obtained, whether it be through a probability sampling design or not, one can still utilize \hat{N}_1 and $\hat{N}_{RB,1}$ to estimate the population size since the Lincoln-Petersen estimator only requires one of the two samples to be obtained completely at random.

4.3 Average node degree estimators

Estimates of the average out-degree of the population members can be useful to the researcher when estimating the characteristics of hard-to-reach populations. We can obtain estimates of the average degree of the population members as follows. For notational purposes, we shall define $d_i = w_i^+$ for all $i = 1, 2, \dots, N$, that is, d_i is equal to unit i 's out-degree. For notational convenience, we shall let $M = \bigcup_{k=1}^K s_{0k}$. We can then estimate the average out-degree of the population, $d_\mu = \frac{\sum_{i=1}^N d_i}{N}$, with the estimator based on the unique members selected for the initial samples, namely

$$\hat{d}_0 = \frac{\sum_{i \in M} d_i}{|M|}.$$

Conditional on $|M|$ this estimator can be viewed as being based on a random sample of $|M|$ individuals selected without replacement. Therefore, conditional on $|M|$, \hat{d}_0 can be shown to be an unbiased estimator for d_μ . The Rao-Blackwellized version of the preliminary estimator of the average node degree is made possible through the same procedure as obtaining the Rao-Blackwellized version of the preliminary estimator of the population size. The corresponding formula used for obtaining the Rao-Blackwellized version of \hat{d}_0 is

$$\hat{d}_{RB} = \sum_{r_1=1}^{n'_1!} \sum_{r_2=1}^{n'_2!} \cdots \sum_{r_K=1}^{n'_K!} \hat{d}_0^{(r_1, r_2, \dots, r_K)} p(s_{(r_1, r_2, \dots, r_K)} | T_S).$$

We shall note here that estimates of the average node degree which are based on

the two-sample study and that are similar to $\hat{N}_{0,1}$ will introduce some bias into the estimator and therefore are not explored in this article.

4.4 Variance estimates

Schwarz and Seber (1999) outlined several methods for obtaining estimates of the variance of capture-recapture estimates based on a K sample study. In our two-sample study we shall take an estimate of the variances of the preliminary estimators \hat{N}_0 and $\hat{N}_{0,1}$ to be the estimator that was proposed by Seber (1970). These estimators are of the form

$$\hat{\text{Var}}(\hat{N}_0) = \frac{(n_{01} + 1)(n_{02} + 1)(n_{01} - m)(n_{02} - m)}{(m + 1)^2(m + 2)}$$

and

$$\hat{\text{Var}}(\hat{N}_{0,1}) = \frac{(n_{01} + 1)(n'_2 + 1)(n_{01} - m_1)(n'_2 - m_1)}{(m_1 + 1)^2(m_1 + 2)}.$$

An estimate of the variance of \hat{d}_0 is the conditionally unbiased estimate

$$\hat{\text{Var}}(\hat{d}_0|M) = \frac{N - |M|}{N} \frac{s^2}{|M|}, \quad (2)$$

where $\frac{N - |M|}{N}$ corresponds with the finite population correction factor and $s^2 =$

$\frac{1}{|M|-1} \sum_{i \in M} (d_i - \hat{d}_0)^2$. As the population size is not known in advance we shall substitute N with \hat{N}_0 or $\hat{N}_{0,1}$ in the finite population correction factor.

To estimate the variance of the improved estimators, Thompson (2006) proposed the following unbiased estimator. For any estimator $\hat{\theta}_{RB} = E[\hat{\theta}_0|T_S]$ for some population unknown θ , where θ_0 is the preliminary estimate, the conditional decomposition of variances gives

$$\text{Var}(\hat{\theta}_{RB}) = \text{Var}(\hat{\theta}_0) - E[\text{Var}(\hat{\theta}_0|T_S)].$$

An unbiased estimator of $\text{Var}(\hat{\theta}_{RB})$ is

$$\hat{\text{Var}}(\hat{\theta}_{RB}) = E[\hat{\text{Var}}(\hat{\theta}_0)|T_S] - \text{Var}(\hat{\theta}_0|T_S).$$

This estimator is the difference of the expectation of the estimated variance of the preliminary estimator over all reorderings of the data and the variance of the preliminary estimator over the reorderings of the data. As this estimator can result in negative estimates of the variance, a conservative approach is take the estimate of $\text{Var}(\hat{\theta}_{RB})$ to be $E[\hat{\text{Var}}(\hat{\theta}_0)|T_S]$ when such a scenario arises.

5 Computational Method for Calculation of Improved Estimators

Due to the large number of sample permutations that are obtained with the sample sizes used in this study, a Markov chain resampling procedure similar to the one found in Thompson (2006) is implemented to obtain estimates of the improved estimates. As the sampling strategy presented in this paper selects multiple independent adaptive web samples, the Markov chain resampling strategy needs to be modified. We outline the modified Markov chain accept/reject (Hastings, 1970) resampling procedure below.

Suppose θ is a population unknown we wish to estimate with the improved estimator $\hat{\theta}_{RB} = E[\hat{\theta}_0 | T_S]$.

Step 0: Let $\hat{\theta}_0^{(0)}$ be the estimated value of θ and $\hat{\text{Var}}(\hat{\theta}_0^{(0)})$ be the estimated value of $\text{Var}(\hat{\theta}_0)$ that is obtained from selecting K adaptive samples in the original order they were selected. Also, let $t^{(0)} = s_{(1,2,\dots,K)}$ be the ordered original samples in the order they were selected.

For step $l = 1, 2, \dots, R$, where R is sufficiently large:

Draw a candidate sample reordering, $t^{(l)}$ say, from a candidate distribution (that is, $t^{(l)}$ is an ordered set of reorderings of each sample). Suppose the most recently accepted candidate reordering is $t^{(y)}$ for some ordered set of reorderings of the samples where $y = 1, 2, \dots, l - 1$. Let $p(t^{(l)})$ be the probability of obtaining $t^{(l)}$ under the true population and $p_q(t^{(l)})$ be the probability of obtaining reordering $t^{(l)}$ under the

candidate distribution. With probability equal to $\min\{\frac{p(t^{(l)})}{p(t^{(y)})} \frac{p_q(t^{(y)})}{p_q(t^{(l)})}, 1\}$, let $\hat{\theta}_0^{(l)}$ and $\hat{\text{Var}}(\hat{\theta}_0^{(l)})$ be the estimates of θ and $\text{Var}(\hat{\theta}_0)$, respectively, obtained with the ordered set of sample reorderings $t^{(l)}$. Otherwise, take $\hat{\theta}_0^{(l)} = \hat{\theta}_0^{(l-1)}$ and $\hat{\text{Var}}(\hat{\theta}_0^{(l)}) = \hat{\text{Var}}(\hat{\theta}_0^{(l-1)})$. Recall that $p(t^{(l)})$ needs only to be known for the (hypothetical) adaptive recruitment probabilities found in the corresponding ordered set of sample reorderings as all terms involving the unknown population size N can be factored out of the ratio of the true probabilities of obtaining sample reorderings and canceled from the expression.

Final step:

Take estimates of $\hat{\theta}_{RB}$ to be

$$\tilde{\theta}_{RB} = \frac{1}{R+1} \sum_{l=0}^R \hat{\theta}_0^{(l)},$$

and similarly take the estimate of $\hat{\text{Var}}(\hat{\theta}_{RB})$ to be

$$\tilde{\text{Var}}(\hat{\theta}_{RB}) = \tilde{E}[\hat{\text{Var}}(\hat{\theta}_0)|T_S] - \tilde{\text{Var}}(\hat{\theta}_0|T_S) = \frac{1}{R+1} \sum_{l=0}^R \hat{\text{Var}}(\hat{\theta}_0^{(l)}) - \frac{1}{R+1} \sum_{l=0}^R (\hat{\theta}_0^{(l)} - \tilde{\theta}_{RB})^2.$$

With the adaptive sampling designs restricted to only recruiting members that are linked to the current active set, and not allowing for random jumps, a large number of the sample reorderings will likely have zero probability of being selected in the full population setting. One primary reason for this is that the sample reorderings that consist of at least one member added after the hypothetical current sample,

with whom do not share a link to any previously selected members that are in the active set, result in a sample that is not sequentially obtainable under an adaptive web sampling design that does not permit for random jumps. Therefore, a candidate distribution which works over each sample individually and first places all corresponding sampled units that are not nominated by any other sampled units into the hypothetical initial sample with probability one is implemented (notice that these members must be in the original initial sample). The candidate distribution then selects the remaining members for each individual sample based on the general adaptive web sampling design, with a small probability of jumps allowed, applied to the reduced population that consists only of those sampled members.

6 Simulation Study

A networked population was simulated for study to evaluate the performance of the sampling designs and inference methods outlined in this article. The population is simulated according to a latent space cluster model and all ties in the population are reciprocated, that is, $w_{ij} = w_{ji}$ for all $i, j = 1, 2, \dots, N$. An illustration of the simulated population can be found in Figure 1. Figure 1 also shows two samples with enlarged graph nodes for ease of visualization that are selected under the general adaptive web sampling design and nearest neighbors adaptive web sampling design where 40 members are selected for the initial samples with (up to) 10 members added adaptively to each sample. The first sample is represented by light colored nodes and the second sample is represented by dark colored nodes. Nodes that are selected for

both samples are highlighted as shaded nodes. Notice the disproportionate increase in the overlap between the adaptively recruited members from the samples for both designs, illustrating the additional information that may be harnessed for inferential purposes.

A simulation study was conducted as follows. A total of 1000 samples with 5000 re-samples from each sample for the Markov chain resampling procedure were obtained with each sampling design. Initial samples of size 40 with (up to) 10 members recruited adaptively were selected for each sample. Histograms of the estimates of the population size and average node degree are shown in Figures 3 and 4, respectively. The true population size of 300 and average node degree of 2.8 are indicated by the solid triangles on the corresponding graphs. All estimates came out approximately unbiased. Table 1 provides the standardized mean squared error (MSE) scores for each of the estimates. The scores are standardized by the MSE score obtained with the preliminary estimates.

As illustrated in the histograms and table of scores, significant gains in precision are made over the preliminary estimates for the population size when using the improved estimates. It also appears that the improved estimates for the average degree of the population made significant gains in precision over their preliminary estimator counterparts. As expected, $\hat{N}_{RB,1}$ has exhibited the best performance, offering some improvement over \hat{N}_{RB} . Further improvement over $\hat{N}_{RB,1}$ can be expected by taking the average of $\hat{N}_{RB,1}$ and $\hat{N}_{RB,2}$ (ie. the Rao-Blackwellized estimator based on taking sample 1 as a fixed list and sample 2 as selected via a link-tracing design). In this study, these estimates came out highly correlated and offered minimal improvement

over $\hat{N}_{RB,1}$.

In our study we did not encounter any negative estimates of the variance of the Rao-Blackwellized estimates and therefore did not have to resort to using the conservative approach that was suggested in Section 4.4. Table 2 gives the coverage rates of the estimates of the population size and average node degree, with average semi-lengths of the 95% confidence intervals in parentheses, based on the Central Limit Theorem. The coverage rates for the population size are smaller than 95%, primarily due to the skewed shape of the distribution of the estimates based on the small sample sizes used in the study. The coverage rates for the average node degree came out close to 95%, indicating that substituting the estimate of the population size into the corresponding variance expression found in (2) is a suitable choice.

7 Discussion

In this article we have outlined a new inferential method that uses link-tracing strategies and a sufficient statistic to estimate the size of hard-to-reach populations. The new method possesses the ability to adaptively recruit hard-to-reach members for the study, through an adaptive sampling probability mechanism that can be tailored to meet the sampler's needs, without introducing additional bias into the improved estimates while allowing for control over sample sizes. As the theoretical results and simulation studies showed, the new methods outlined in this article will give rise to more precise estimators relative to those hitherto found in the existing capture-recapture literature.

One additional advantage the new methods presented in this article possess over some of the existing capture-recapture methods is outlined as follows. In some empirical settings when sampling from a large population with relatively small sample sizes, the selection of two random samples may give rise to little or no overlap in the samples, hence rendering an undesirable estimate of the population size when using a capture-recapture style of estimator. With the methods outlined in this article, overlap between the adaptive recruitment stages of the samples is more certain and hence the use of the new inferential procedure should result in a much more reliable estimate of the population size.

The minimal sufficient statistic, as described in Thompson and Seber (1996), was not used in this study primarily for computational reasons. The use of the minimal sufficient statistic is deserving of attention and should be explored in future work as gains in improvements of the estimates over those found with the sufficient statistic used in this article are certain.

Extending on the methods outlined in this article to be compatible with the eight closed population models commonly used in capture-recapture studies, namely the $M_0, M_b, M_t, M_h, M_{tb}, M_{bh}, M_{th}, M_{tbh}$ models (Schwarz and Seber, 1999), is certainly deserving of future attention. This should be straightforward as the corresponding inference procedures work over random sample sizes, and conditioning on the observed sample sizes will permit for Rao-Blackwellization of the corresponding capture-recapture estimates.

8 Acknowledgements

The authors wish to thank the Natural Sciences and Engineering Council for supporting this work.

9 References

- Abdul-Quader, A., Heckathorn, D., McKnight, C., Bramson, H., Nemeth, C. Sabin, K., Gallagher, K., and Des Jarlais, D. (2006). Effectiveness of respondent-driven sampling for recruiting drug users in New York City: Finding from a pilot study. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, **83** (3) 459-476.
- Chapman, D. (1951). Some properties of the hypergeometric distribution with application zoological census. *Univ. Cal. Pub. Stat.*, **1** 131-160.
- Chao, A., Tsay, P., Lin, S., Shau, W., and Chao, D. (2001). Tutorial in Biostatistics: The application of capture-recapture models to epidemiological data. *Statistics in Medicine* **20** (3) 3123-3157.
- Chow, M. and Thompson, S. (2003). Estimation with link-tracing sampling designs-a Bayesian approach. *Survey Methodology* **29** (2) 197-205.
- Feinberg, S. (2010a). Introduction to papers on the modeling and analysis of network data. *The Annals of Applied Statistics*, **4** (1) 1-4.
- Feinberg, S. (2010b). Introduction to papers on the modeling and analysis of network data-II. *The Annals of Applied Statistics*, **4** (2) 533-534.
- Felix-Medina, M. and Thompson, S. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics*, **20** (1) 19-38.

- Felix-Medina, M. and Monjardin, P. (2006). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations: A Bayesian-assisted approach. *Survey Methodology*, **32** (2) 187-195.
- Frank, O. and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, **10** (1) 53-67.
- Frischer, M., Leyland, A., Cormack, R., Goldberg, D., Bloor, M., Green, S., Taylor, A., Covell, R., McKeganey, N. and Platt, S. (1993). Estimating the population prevalence of injection drug use and infection with Human Immunodeficiency Virus among injection drug users in Glasgow, Scotland. *American Journal of Epidemiology*, **138** (3) 170-181.
- Gile, K. and Handcock, M. (2011). Network model-assisted inference from respondent-driven sampling data. Arxiv preprint arXiv:1108.0298 .
- Handcock, M. and Giles, K. (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics*, **4** (1) 5-25.
- Hastings, W. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika*, **57**, 97-109.
- Heckathorn, D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, **44** (2) 174-199.
- Heckathorn, D. (2002). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, **49** (1) 11-34.

Mastro, T., Kitayaporn, D., Weniger, B., Vanichseni, S., Laosunthorn, V., Uneklabh, T., Uneklabh, C., Choopanya, K., and Limpakarnjanarat, K. (1994). Estimating the number of HIV-infected injection drug users in Bangkok: A capture-recapture method.

Petersen, C. (1896). The yearly immigration of young Plaice into the limfjord from the German Sea. *Report of the Danish Biological Station*, **6** 5-84.

Schwarz, C. and Seber, G. (1999) Estimating animal abundance: Review III. *Statistical Science*, **14** (4) 427-456.

Seber, G. (1970). The effects of trap response on tag-recapture estimates. *Biometrika* **26**, 13-22.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why?. *Bulletin de Methodologie Sociologique* **36**, 34-58.

Thompson, S. and Seber, G. (1996). *Adaptive Sampling*. Wiley.

Thompson, S. and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, **26**, (1) 87-98.

Thompson, S. (2006). Adaptive web sampling. *Biometrics*, **62** 1224-1234.

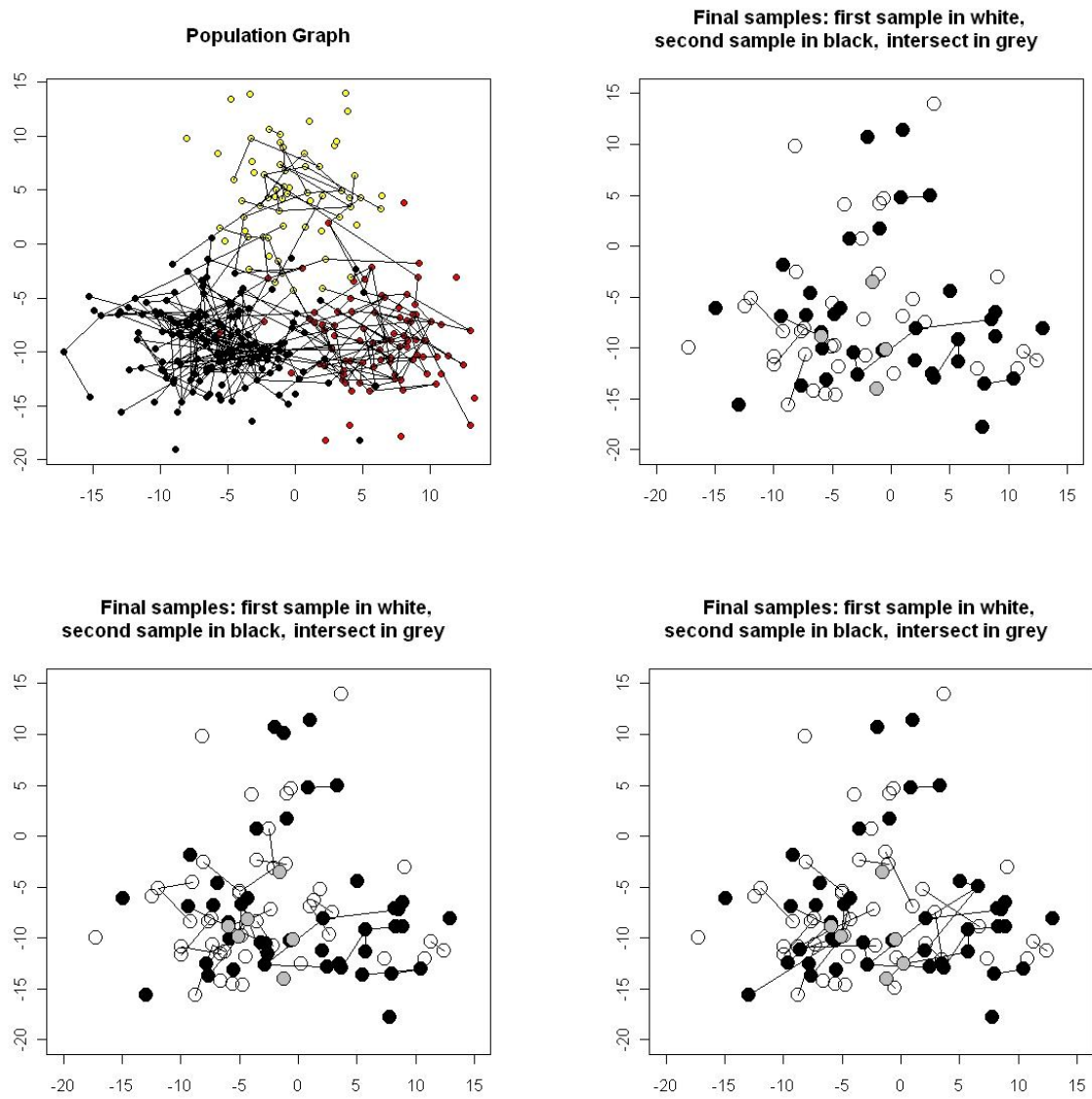


Figure 1: The simulated study population on the top left. Two random initial samples on the top right with two general adaptive web samples on the bottom left and two nearest neighbors adaptive web samples on the bottom right.

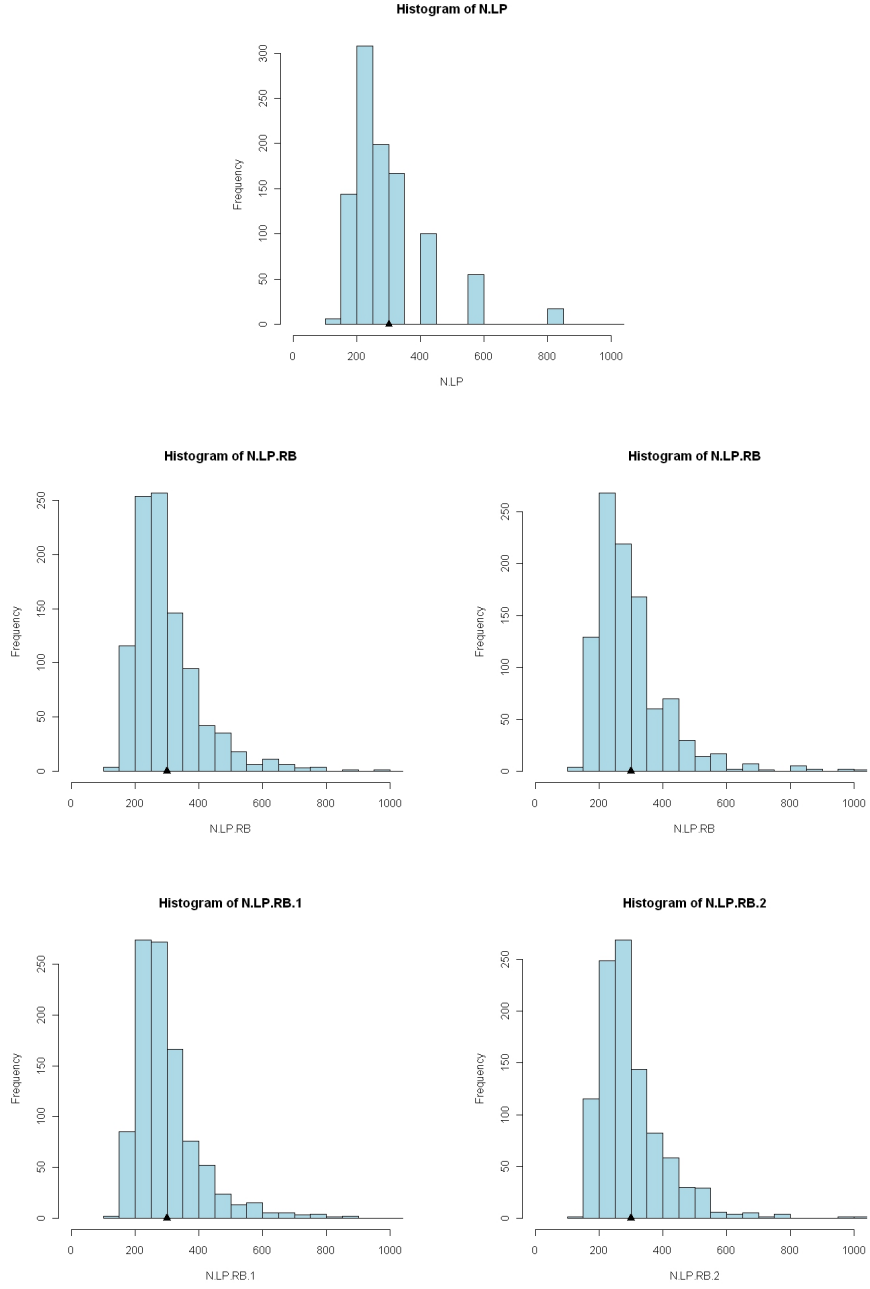


Figure 2: Histogram of \hat{N}_0 on top and \hat{N}_{RB} based on the general adaptive web sampling design and the nearest neighbors adaptive web sampling design, respectively, in the middle. Histograms of $\hat{N}_{RB,1}$ based on the general web adaptive sampling design and the nearest neighbors adaptive web sampling design, respectively, on the bottom.

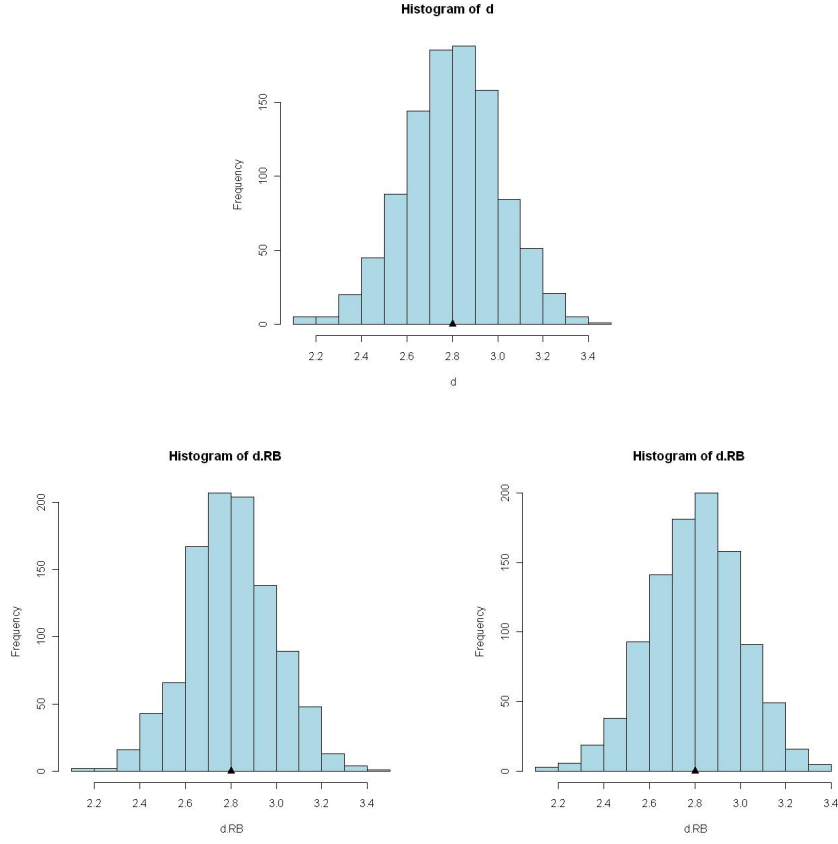


Figure 3: Histograms of \hat{d}_0 on top and \hat{d}_{RB} on the bottom based on the general adaptive web sampling design and the nearest neighbors adaptive web sampling design, respectively.

Table 1: Standardized MSE scores for the estimates.					
Estimator	\hat{N}_0	\hat{N}_{RB} , Gen.	\hat{N}_{RB} , NN	$\hat{N}_{RB,1}$, Gen.	$\hat{N}_{RB,1}$, NN
Parameter					
N	1	0.566	0.637	0.518	0.594
d_μ	1	0.857	0.947		

Table 2: Coverage rates of the estimates with average semi-length of 95% confidence intervals in parentheses.

Estimator	\hat{N}_0	\hat{N}_{RB} , Gen.	\hat{N}_{RB} , NN	$\hat{N}_{RB,1}$, Gen.	$\hat{N}_{RB,1}$, NN
Parameter					
N	0.850 (211)	0.854 (178)	0.851 (189)	0.868 (169)	0.860 (177)
d_μ	0.938 (0.401)	0.931 (0.363)	0.922 (0.378)		